Performance Analysis of Event Detection Models in Crowded Scenes

Ernesto L. Andrade[1], Scott S. Blunsden[2] and Robert B. Fisher[1]
School of Informatics, University of Edinburgh
King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, UK
[1] {eaneto,rbf}@inf.ed.ac.uk, [2]S.J.Blunsden@sms.ed.ac.uk

Keywords: surveillance, crowd, optical flow, HMM, learning

Abstract

This paper evaluates an automatic technique for detection of abnormal events in crowds. Crowd behaviour is difficult to predict and might not be easily semantically translated. Moreover it is difficult to track individuals in the crowd using state of the art tracking algorithms. Therefore we characterise crowd behaviour by observing the crowd optical flow and use unsupervised feature extraction to encode normal crowd behaviour. The unsupervised feature extraction applies spectral clustering to find the optimal number of models to represent normal motion patterns. The motion models are HMMs to cope with the variable number of motion samples that might be present in each observation window. The results on simulated crowds analyse the robustness of the approach for detecting crowd emergency scenarios observing the crowd at local and global levels. The results on normal real data show the effectiveness in modelling the more diverse behaviour present in normal crowds. These results improve our previous work [1] in the detection of anomalies in pedestrian data.

1 Introduction

In recent years computer vision and machine learning techniques have been applied to modeling and recognition of human activities and interactions. The application domains for these techniques usually involve simple environments such as offices [9], kitchens [4], cargo bays [7] and loading docks [6] such that activity recognition is focused upon modeling the actions and interactions of small groups of people/objects. However, there have been a few attempts to model larger groups of people, crowds, which are mostly based on discriminative classifiers [13]. The analysis of crowd movements and behaviour is of particular interest in surveillance domain [8]. In scenarios where hundreds of cameras are monitored by a few operators behavioural analysis of crowds is useful as a tool for video pre-screening.

In order to model a crowd the model must cope with a large variation in densities and motions present in a real crowd. This requires a huge amount of data to enable a good supervised/unsupervised learning for discriminative or generative crowd models. Moreover in the surveillance domain usually there are few or no examples of the emergency/abnormal events to be detected. Thus the first assumption for our crowd modelling is that we are trying to model the degree of similarity between the trained model and the new unseen video data. Therefore the events are classified as normal or abnormal behaviour without having any other particular labels for them. This arrives from the fact that crowds are difficult to treat semantically. In a real crowd scene one can not beforehand easily specify or train particular labels for behavioural analysis. This scene content labelling would discretise the input space simplifying the analysis. However, unsupervised learning techniques provide the means to learn the typical labels (space-time behavioural patterns) and have been applied for similar problems in video analysis [18] [14]. In our work the analysis is based on the optical flow patterns of scene. To reduce dimensionality we project the input optical flow patterns on the principal components of the training flow fields. This compressed feature set is used by learning algorithms. The automatic model extraction involves fitting an HMM for each video segment and performing spectral clustering on the similarity matrix computed using inter-segment likelihoods. The resulting clustered video segments are used to train a new set of HMMs which represent the optical flow variations on the normal example set. Abnormality detection is based on a threshold on the HMM bank likelihood function. This framework is applied to detect simulated emergencies in crowds. In addition we show an example of the same technique applied to real data. However, for the real scenes no emergencies are present and this data is used to illustrate the framework modelling capabilities and the lack of false positives.

This paper is organised as follows section 2 discusses the related research. Section 3 details the methodology for anomaly characterisation and detection. Section 4 presents results on simulated and real data. Section 5 draws the final comments and conclusions.

2 Related Work

The use of principal component analysis of optical flow fields as features is demonstrated in [5], where principal components of video sequences are used to construct a linear basis for complex motion phenomena. Unusual events are analysed in a similar context in [7] and [17] where deviations from example normal behaviour are used to characterise abnormality. Spectral clustering using HMMs as similarity measures is used

for trajectory classification in [10]. In another related spectral clustering application it is used to automatically determine models for video sequence in [14]. Our approach is based on the general concepts in these references and to the best of our knowledge our work is the first combined application of the techniques of optical flow, subspaces and HMMs to assess similarity to the problem of abnormal behaviour detection in crowds. This builds up on our previous work in [1] where similar ideas of optical flow similarity based on HMMs were used to analyse variations in the flow patterns of pedestrian traffic. The current work allows for a more flexible model which is in principle able to deal with a large range of people density in the scene from sparse pedestrian traffic to dense crowd flows.

3 Methodology

The characterisation of normal behaviour for the crowd uses normal optical flow patterns to estimate the model parameters. The modelling process involves four phases: 1) Preprocessing: background modelling and optical flow computation; 2) Feature prototypes: principal components analysis on the example flow fields, 3) Spectral Clustering: automatic determination of the number of HMMs to represent the flow sequences and 4) Bank of models: training of the HMM models using the data of each cluster per model. The analysis concentrates on identifying unusual events in the crowd by comparing the new observation's likelihood to a detection threshold. Details of this are given in the next subsections.

Preprocessing involves the construction of a Mixture of Gaussians background model for the scene based on [12]. The background model produces a mask with the detected foreground objects per frame. In parallel to foreground extraction robust optical flow is computed for the whole frame using the techniques described in [3]. Prior to the optical flow computation the sequence is smoothed with a 5x5x5 Gaussian filter ($\sigma = 0.8$) to reduce acquisition noise. The resulting optical flow is sub-sampled by median filter with a window of size 8x8 applied independently to the horizontal and vertical components. The combination of flow information with the foreground mask allows the analysis to only consider flow vectors inside foreground objects, reducing observation noise. The motion parameters are encoded in a sample vector of the form $\mathbf{s} = (u, v)$, where u and v are horizontal and vertical optical flow components. Prior to the analysis the foreground mask is superimposed on the optical flow output resulting in the motion parameters for the detected foreground objects. All values outside the foreground mask are set to zero to characterise the static regions.

3.1 Video Segmentation

The assumption for video segmentation is that there is no distinctive activity or periods of inactivity everywhere in the training crowd video. Therefore all segments in the video stream are equally important for prototype extraction. The video sequence $\mathbf{V}$ is segmented into N video segments $\mathbf{V} = \{\mathbf{v_1}, ..., \mathbf{v_N}\}$ of equal length T frames, $\mathbf{v_n} = \{\mathbf{F_{n1}}, ..., \mathbf{F_{nT}}\}$ as in [18], $\mathbf{F_{nt}} = (s_1, ..., s_P)$, where P is the number of flow vectors in each frame. $T = 100$ frames (4 seconds) is assumed in the experimental section to contain enough crowd movement for comparison.

3.2 Feature Prototypes

The first step of the prototype extraction is to perform principal component analysis (PCA) on the optical flow fields of each frame $\mathbf{F_{nt}} = ((u_1, v_1), ..., (u_P, v_P))$ in $\mathbf{V}$. The first J eigenvectors with the largest eigenvalues are selected to form a basis for the projection. The projection reduces the input feature dimensionality from the dimension of flow fields samples $2 \times P$ to the dimension of the selected eigenvectors J. The resulting set of feature vectors for the n$-th$ segment in $\mathbf{V}$ is:

$$\mathbf{W_n} = \{\mathbf{w_{n1}}, ..., \mathbf{w_{nT}}\} \quad (1)$$

where $\mathbf{w_{nt}}$ is a vector representing the projection of the t-th frame in the n-th segment over the selected eigenvectors, defined as

$$\mathbf{w_{nt}} = \{g_{nt1}, ..., g_{ntJ}\} \quad (2)$$

where g_{ntm} is the weight associated with the m-th eigenvector.

3.3 Spectral Clustering

The derivation of the similarity measure of the video segments for spectral clustering is based on likelihood of the observations in the segments given by a Hidden Markov model. For that a Mixture of Gaussian Hidden Markov Model (MOGHMM) [11] is trained with the feature vectors for each video segment inside the training set resulting in $\mathbf{B_n}, n = 1..N$ models (ie. one for each segment). We use the MOGHMM to model the different patterns of optical flow inthe image. This MOGHMM structure is ergodic with J states (same as the number of selected eigenvectors) and M Gaussian emission probabilities per state with diagonal covariance matrices in order to reduce the number of samples needed to train the MOGHMM (assuming independence in the input space of

eigenvectors projections). In the text below $L(\ )$ is the log-likelihood of the model defined as the sum of the logarithm of the scaling factors in the forward-backward procedure [11].

The measure of similarity between video segments is defined as:

$$S_{ij} = \frac{1}{2}\left\{L(\mathbf{W_j}|\mathbf{B_i}) + L(\mathbf{W_i}|\mathbf{B_j})\right\} \quad (3)$$

The pairwise similarity values between the video segments forms a similarity matrix $\mathbf{S}$. The similarity matrix is subject to spectral clustering using the algorithm described in [15] to automatically find the number of groups in the video data. We use this clustering method because it automatically selects the number of natural clusters in the dataset using a local scaling strategy, other methods of clustering may work if you could estimate the number of clusters.

3.4 HMM Training

After spectral clustering the video segments are regrouped into a more compact number of classes K. All the samples $\mathbf{W_n}$ in each class are used to train a new MOGHMM per class $\mathbf{M}_k$. The final model for the video sequence has the form:

$$L(\mathbf{W}|\mathbf{M}) = max_k(L(\mathbf{W}|\mathbf{M}_k)) \quad (4)$$

where $\mathbf{W}$ are the samples in the model bank observation windows.

3.5 Event Classification

The classification of normal and abnormal events is based on the comparison of the current observation's likelihood given by the bank of MOGHMM models and the detection threshold. The observation of the n-th test video segment $\mathbf{W_n^o}$ (the superscript $\mathbf{o}$ denotes new observations not used in training) is considered abnormal if:

$$L(\mathbf{W_n^o}|\mathbf{M}) < Th_{Ab} \quad (5)$$

The test video features $\mathbf{W_n^o}$ are extracted by projecting the test flow fields on the J eigenvectors of the sub-space derived from the training set.

3.6 Local Analysis

The previous subsections approach applied the analysis framework to the whole frame and therefore this analysis is called global.To detect small variations (which can be hidden in the likelihood function oscillations of the globalmodel) we describe the application of the same framework to small areas of the original frame which we call local analysis. IN this analysis the original optical flow frame is divided in non-overlapping patches. For each one of the selected patches of the original frame of width bw and height bh the same subspace analysis is performed now taking only the flow vectors in the training video set inside the patch to compose a local basis. This local basis is used in the same manner as in the global analysis now producing a local set of MOGHMM models which encode more specifically the optical flow variations inside the patch. Abnormalities are checked in the same way by comparing the deviations of the normal local model against a local detection threshold, which can be adapted per frame patch.

The local model is applied to all blocks in the flow field. To allow for on-line event detection the likelihood drops are measured with a simple edge filter on the likelihood function. Long lasting likelihood drops within the filter indicate the abnormal events. The filter delays are adjusted to provide the desired false alarm rate. The detection filter equation is:

$$F_t(L) = |\frac{\sum_{l=t-W_s/2}^{t} L(l)}{W_s/2+1} - \frac{\sum_{l=t+1}^{t+W_s/2} L(l)}{W_s/2}| \quad (6)$$

where t is the current frame, $W_s = 200$ (8 secs) is the observation window and $L(l)$ is the model log-likelihood for the l-th frame.

4 Experimental Results

4.1 Simulated Crowd Data

There are three simulated data sets: normal flow, blocked exit and person dropping on the floor. In the normal flow simulation a crowd flows in one direction in the scene. In the blocked exit simulation the crowd cannot leave the scene and starts to press against the exit. In the person dropping on the floor scenario when the person falls in the middle of the crowd the others start to deviate to avoid steepping over the fallen person. The simulation technique is described in [2]. The original frame size is 384x288 pixels and the optical flow observations are subsampled, by the u and v median over 8x8 blocks, resulting in optical flow image of 48 x 36 (P=1728) flow vectors. One normal simulated sequence with 2000 frames is used for training. It is divided for clustering in $N = 20$

segments of size $T = 100$. $K = 13$ video segments clusters is the mean number of clusters automatically selected by the spectral clustering algorithm in 30 runs. For the test there are 10 simulations of the blocked exit event.

In the blocked exit scenario we have evaluated different HMM topologies looking for the largest mean drop in likelihood over the 10 simulated sequences as the criteria to choose the best model topology for this emergency scenario. For all the topologies with different number of states (Q) and input features (J) the number of Gaussians per state (M) is constant and experimentally determined to be 3 using Mixture of Gaussians fitting in the distributions of the eigenvectors projections. These results are summarised in table 1 which presents the likelihood drop after the event as a function of Q and J. Tables 2 and 3 present the variations of the likelihood standard deviation before and after the event respectively. The drop is highest for the topology with $Q = 10$ states and $J = 10$ features. We have chosen this topology for the detector test although it was observed during training that as we add more feature/eigenvectors we can obtain other models with similar performance but requiring 4 to 5 times more features to present the same drops. One of the eigenvectors of the simulated normal optical flow fields used for training/feature extraction is shown in Fig. 1, where we notice the regularity of the unidirectional flow in the simulation.

The results for the detection of the blocked exit event for the 10 simulation runs are shown in Fig.2. There is a clear and quick drop in the likelihood function less than 100 frames (4 seconds) after the exit blocking. A threshold Th_{Ab} slightly larger than three standard deviation of the normal flow (i.e $Th_{Ab} = 3 \times 6$) would guarantee the detection of this event within less than 200 frames with no false alarms. The size of the observation window used to compute the likelihood in Fig. 2 is 50 frames. Larger window sizes tend to smooth the likelihood function reducing the sensitivity of the detector.

For the more subtle perturbation in the flow of the person drop scenario we apply the local analysis by aggregating the original flow field of size 48 x 36 flow vectors in blocks of $bw = 4$ x $bh = 4$. This results in 108 blocks each having its own model fitting and training procedure similar to the global analysis described before See 3. We show here the results for the person drop event only in the blocks close to where the event occurred in the scene. The other blocks do not show any significant deviation in likelihood after the event and are not shown here due to space constraints. The topology investigation for the models in the area of the event is similar to the global analysis and for that the maximum drop criteria has selected an MOGHMM with $Q = 10$ states and $J = 20$ eigenvectors with spectral clustering determing the size of models in the local MOGHMM to be $K = 6$. Table 4 summarises the mean variations before and after the event although there is a drop in the likelihood this variations is obscured by the likelihood response oscillations (see Table 5). This justifies the use of the filter $F_t(L)$ to detect with a delay ($Ws = 200$ frames,8 sec) the moment where the drop in likelihood is more intense and use this event as the detection criteria for the person fall in the crowd.

Fig. 4 shows the statistics of the temporal edge filter for all 10 runs of the person fall event. The filter is applied to the likelihood response of each block around the area where the person falls. The only noticeable increases in the response are at the blocks close to the person falling and no other detections above such levels are present in the other blocks through the whole sequence, where people continue to move in their normal way. Fig. 4 shows that this local emergency is close to the detection limit for the flow based approach and altought we are able to detect the flow alteration the detection is not so reliable as for the global case mainly due to two factors: i) the event scale (eg. about less than 1 block in size) is at the resolution limit of the flow field observations and there are still interferences of the surrounding flows (eg. people passing close to the fallen person in their normal directions).

Figure 1: Eigenflows for the simulated normal training set (elements in the first eigenvector).

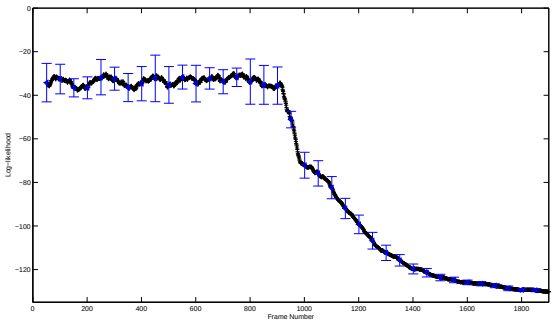

Figure 2: Log-likelihood results for the blocked exit event. Normal flow before frame 900, blocked exit after frame 900. $Q = 10$ states, $M = 3$ gaussian per state, $J = 10$ eigenvectors and $K = 13$ models. Error bars show one standard deviation.

4.2 Real Data

In order to illustrate the model applicability to real data two real sequence of 5000 frames of normal crowd motion are submitted to the global analysis framework. One is used to

Q / J	1	10	20	30	40	50
2	-5.04	-36.10	-36.72	-34.00	-38.58	-47.38
3	-6.29	-45.97	-45.03	-38.39	-25.83	-54.75
4	-6.91	-49.77	-47.16	-47.14	-53.22	-60.50
5	-6.30	-79.79	-43.69	-40.29	-62.57	-56.73
6	-7.54	-60.68	-53.77	-49.61	-71.49	-62.37
10	-7.73	$\mathbf{-80.38}$	-62.75	-57.50	-73.56	-77.66

Table 1: Drop of the mean likelihood from before to after block event versus number of states (Q) and number of input feature eigenvectors(J). $M = 3$ gaussians per state and $K = 13$ in the model bank after training.

Q / J	1	10	20	30	40	50
2	1.82	1.59	2.59	2.90	3.61	5.38
3	1.83	1.85	3.25	4.07	2.19	7.36
4	2.87	1.92	3.24	5.52	6.36	8.26
5	2.95	2.99	3.76	5.59	8.23	8.15
6	3.16	2.38	4.69	6.89	9.87	9.53
10	4.64	4.56	6.75	9.48	11.26	13.55

Table 2: Mean standard deviation before block event versus number of states (Q) and number of input feature eigenvectors(J). $M = 3$ gaussians using max model output. $K = 13$ models.

Q / J	1	10	20	30	40	50
2	13.33	15.46	16.20	15.05	17.08	20.33
3	16.04	19.82	19.47	17.34	10.92	24.09
4	22.49	21.42	19.94	22.06	25.19	26.70
5	19.99	32.83	18.98	18.88	31.06	25.37
6	19.62	25.08	23.28	23.62	35.66	28.45
10	30.72	32.64	28.88	28.16	36.95	34.47

Table 3: Mean standard deviation after block versus number of states (Q) and number of input feature eigenvectors(J). $M = 3$ gaussians using max model output. $K = 13$ models.

		Event Type	
Interval	Block Position	normal	drop
Before	Left	16.88	17.04
	Event	15.00	**15.47**
	Right	15.19	15.28
After	Left	17.92	16.84
	Event	16.18	**10.05**
	Right	15.84	13.74

Table 4: Local analysis mean likelihood for $Q = 10$ states, $J = 20$ eigenvectors, $M = 3$ gaussians and $K = 6$ models.

		Event Type	
Interval	Block Position	normal	drop
Before	Left	2.69	3.01
	Event	11.63	**8.64**
	Right	6.33	5.65
After	Left	9.23	6.79
	Event	20.56	**26.51**
	Right	20.32	19.58

Table 5: Local analysis standard deviation of likelihood for $Q = 10$ states, $J = 20$ eigenvectors, $M = 3$ gaussians and $K = 6$ models.

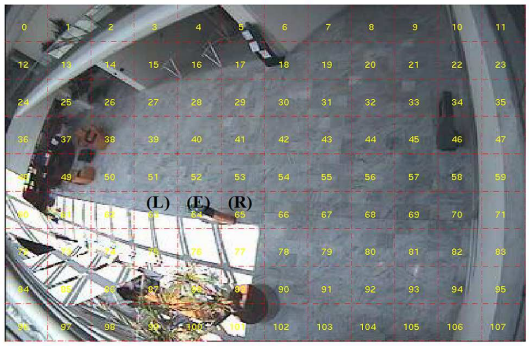

Figure 3: Local emergency detection. (L) and (R) blocks to the left and right of the event respectively and (E) block where the event occurs.

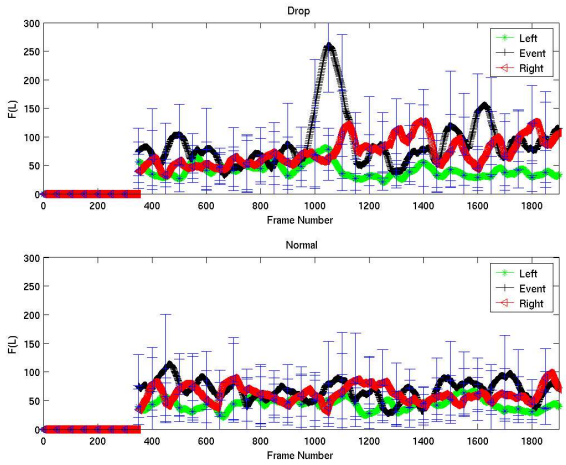

Figure 4: Detection results for the local analysis. Top, drop person event. Bottom, normal flow. Error bars show one standard deviation.

train the model and the other one is evaluated against it. The scene contains people moving about in different directions in a train station. The first eigenvector of the optical flow sequence is shown in Fig. 5, showing the most frequent flow directions in the scene. The results for the trained bank of models likelihood applied to the test sequence are shown in Fig. 6 where the normal crowd motion is encoded with a bank of models with $Q = 10$ states, $M = 3$ gaussians per state, and $J = 30$ eigenvectors resulting in $K = 14$ models. The likelihood has a larger standard deviation ($Mean = 21.377$,$Std.Dev. = 9.51$) when compared to the normal situation for the simulated crowd data and exemplifies the complexity of event detection in real data, i.e. emergencies to be detected should be outside the range of the normal model fluctuations.

5 Conclusion

This work demonstrated a novel technique for automatic detection of abnormal events in crowds. Using projections

Figure 5: Eigenflows for the real train station data (elements in the first eigenvector).

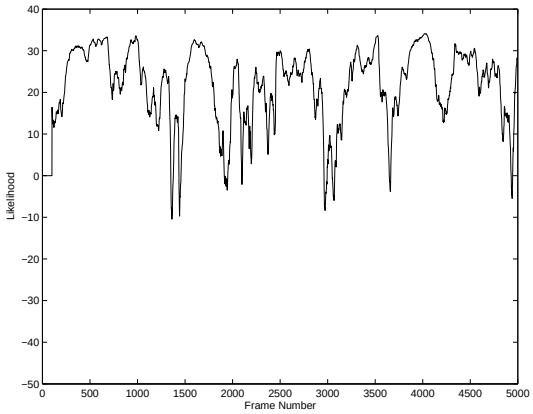

Figure 6: Log-likelihood results for the real scene data in the train station ($Q = 10$ states, $J = 30$ eigenvectors, $M = 3$ gaussians and $K = 14$ models).

of the eigenvectors in a sub-space spanned by the normal crowd scene as an input feature the proposed technique applies spectral clustering to automatically identify the number of distinct motion segments in the sequence. The features in the clustered motion segments are used to train different MOGHMMs for the normal sequence, which compose a bank of models for the training simulated video. The experiments show that the bank of models is effective in quickly detecting the simulated emergency situation in a dense crowd (eg. less than 200 frames after the event for the blocked exit scenario). In the fallen person scenario the drop in the likelihood was well characterised by the local model, however it presented a high variance before and after the event detection requiring additional filtering to provide a reliable detection. The investigation of the relation between the number of eigenvectors, HMM topology and the model likelihood variations before and after the event indicates that optimal configurations should be tested to provide more reliable results for a particular detection task. Our method adds to the detection techniques in [13] allowing for a different representation of the flow dynamics in the crowd.

References

[1] E. L. Andrade, S. J. Blunsden, and R. B. Fisher. Characterisation of optical fbw anomalies in pedestrian traffi c. *The IEE International Symposium on Imaging for Crime Prevention and Detection*, pages 73–78, 2005.

[2] E. L. Andrade and R. B. Fisher. Simulation of crowd problems for computer vision. *First International Workshop on Crowd Simulation*, (3):71–80, 2005.

[3] M. J. Black and P. Anandan. A framework for the robust estimation of optical fbw. *4^{th} International Conference on Computer Vision*, pages 231–236, 1993.

[4] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 1:838–845, 2005.

[5] D. J. Fleet, M. J. Black, Y. Yacoob, and A. D. Jepson. Design and use of linear models for image motion analysis. *International Journal of Computer Vision*, 36(3):171–193, 2000.

[6] S. Gong and T. Xiang. Recognition of group activities using a dynamic probabilistic network. *Proceedings of the IEEE International Conference on Computer Vision*, pages 742–749, 2003.

[7] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman. Detection and explanation of anomalous activities: representing activities as bags of event n-grams. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 1:1031–1038, 2005.

[8] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviours. *IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews*, 43:334–352, 2004.

[9] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, 96:163–180, 2004.

[10] F. Porikli. Learning object trajectory patterns by spectral clustering. *Proceedings IEEE International Conference on Multimedia and EXPO (ICME)*, pages 1171–1174, 2004.

[11] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[12] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[13] S. A. Velastin, B. A. Boghossian, and A. Lazzarato. Detection of potentially dangerous situations involving crowds using image processing. *Proceedings of the Third ICSC Symposia on Intelligent Industrial Automation (IIA'99) and Soft Computing*, 1999.

[14] T. Xiang and S. Gong. Video behaviour profi ling and abnormality detection without manual labelling. *Proceedings IEEE International Conference on Computer Vision*, 2005.

[15] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. *Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, MA*, 2005.

[16] B. Zhan, P. Remagnino, and S.A. Velastin. Mining paths of complex crowd scenes. *Advances in Visual Computing: First International Symposium*, pages 126–133, 2005.

[17] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Semi-supervised adapted hmms for unusual event detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 1:611–618, 2005.

[18] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, 2:819–826, 2004.